Ovarian Cancer Diagnosis Using Discrete Wavelet Transform Based Feature Extraction from Serum Proteomic Patterns

H. Montazery Kordy[1], M. H. Miranbaygi[1], M. H. Moradi[2]
[1]Department of Electrical Engineering, Tarbiat Modarres University, Tehran, Iran
[2]Department of Biomedical Engineering, Amirkabir University of Technology, Tehran, Iran
E-mail: hmontazery@modares.ac.ir

***Abstract*—Pathological changes within an organ might be reflected in proteomic patterns in serum. Mass spectrometry is becoming an important tool that generates the proteomic Patterns. Mass spectrometry yields complex functional data for which the features of scientific interest are the peaks. Due to this complexity of data, a higher order analysis such as wavelet transform is needed to uncover the differences in proteomic patterns. We have applied wavelet based feature extraction method to available data and used a filter approach to feature subset selection in order to identify the appropriate biomarkers from reconstructed mass spectra. Using different classification algorithms, our approach yielded an accuracy of 98%, specificity of 97%, and sensitivity of 100%.**

Keywords: **Proteomics, Cancer diagnosis, Wavelet transform, Classification, Biomarker.**

I. INTRODUCTION

THE development of tools for the early cancer diagnosis has shown to be a major problem, and clinicians have investigated a variety of diagnosis techniques. Recently, it has been found that pathological changes within an organ might be reflected in proteomic patterns in serum [1]. The word 'proteome' coined in 1994, designates the complete set of proteins that ultimately results from genome transcription in a given cell, tissue, or organism [2]. Hence, Proteomics is the science of making qualitative and quantitative comparisons and differentiation among proteomes under various conditions (normal vs. cancer, treated vs. untreated) to understand biological processes (disease). The field of proteomics has since evolved to include almost any type of technology that focuses upon the wide-scale analysis of proteins [3][4].

The mass spectrometry is a tool which provides information about proteins and their fragments. The definition of a mass spectrometer may seem simple: it is an instrument that can ionize a sample and measure the mass-to-charge ratio of the resulting ions. Therefore, mass spectrometer can give qualitative and quantitative information on the elemental, isotopic, and molecular composition of organic samples [5]. The mass spectrum analysis is a fast inexpensive procedure based on a sample of patient's blood, and it may potentially allow cancer screening without any complication in time of diagnosis.

Application of mass spectrometry for diagnosis of ovarian cancer could have an important effect on public health, but to achieve this goal new biomarkers are essential. For women with high risk of ovarian cancer due to family or personal history of cancer, there are no effective screening options. Ovarian cancer presents itself at a late clinical stage in more than 80% of patients, and 35% of this population survive 5-year after diagnosis [6]. On the other hand, the 5-year survival for patients with stage I ovarian cancer exceeds 90%, and increasing the number of women diagnosed with stage I should have a direct effect on the mortality and morbidity of disease.

From a modeling viewpoint, the mass spectra can be considered complex functional data in which the key features of scientific interest are the peaks in the mass spectrum curve [7]. The peaks represent proteins or protein fragments (peptides) in proteomic pattern. Raw mass spectrometry data tends to be incomplete, noisy, highly correlated within the spectrum profile, highly dimensional, etc. and hence not directly suitable for feature extraction. Additionally, mass spectral data display variations in the protein profile even at identical instrumental settings and sample conditions. In order to minimize the effect of irrelevant sources of variations such as humidity, time, etc. and to be able to extract the information of mass spectra in more details, a more sophisticated preprocessing method that de-noises as well as compresses the data needs to be utilized.

The wavelet transform (WT) is an effective tool for dimension reduction and noise removal in the analysis of proteomic data. Wavelets are very popular in signal processing because they are able to analyze both local and global behavior of functions. The WT is a projection of the spectrum onto an orthogonal basis, called a wavelet basis [8]. This is to say that the spectrum can be represented by a set of localized orthogonal basis functions called wavelets. Thus, wavelet analysis could provide de-noised and compressed representation of mass spectrometry data that make the feature extraction process more efficient and accurate due to many favorable properties, such as hierarchical and multiresolution decomposition structure, de-correlated coefficients, and a wide variety of orthogonal basis function possibilities.

We have applied wavelet-based feature extraction method to the mass spectra of ovarian cancer patients and those of healthy people. We have used a filter approach for feature subset selection. We have employed the reconstr-ucted mass spectra to identify the appropriate biomarkers and to evaluate the classification performance. Our results have confirmed that the mass spectrometry proteomic profiles allow the diagnosis of ovarian cancer. Therefore, the wavelet-based reconstructed mass spectra can be a viable method in diagnosis of ovarian cancer. For our developed technique, the accuracy was 98% on the data sets, its specificity was 97%, and its sensitivity was 100%.

II. Data and Preprocessing

We have used the surface enhanced laser desorption-ionization time-of-flight (SELDI-TOF) mass spectrometry (MS) serum proteomic patterns as the input data to our processing algorithms. Serum SELDI-TOF mass spectra data were then used for screening patients and a healthy population.

A. Dataset

Two serum SELDI-TOF MS data sets were used in this research to identify serum proteomic patterns that differentiate the serum of ovarian cancer from non-cancer controls cases. The data set were downloaded from the freely available datasets of the American National Cancer Institute (NIC). As explained on the website, Dataset I was collected using the WCX2 protein chip, and includes 216 samples which has been divide into 100 controls, 100 ovarian cancer, and 16 benign. Dataset II was also collected using WCX2 protein array, but a new set of ovarian samples was used. The sample set included 91 controls and 162 ovarian cancers.

A mass spectrum is a curve where the x-axis indicates the ratio of the weight of a specific molecule to its electrical charge (M/Z, in Daltons per unit charge) and the y-axis is the signal intensity for the same molecule as a measure of the abundance of that molecule in the sample. Each mass spectrum curve represents the expression profile of 15154 peptides defined by their M/Z ratios with corresponding intensities. The distribution of samples for two datasets is represented in Table I. All samples in each dataset are divided into cancer and non-cancer (including control and benign cases) classes in this study. Criticisms have been raised over the validity of the datasets used in this research [9]. Fig. 1 shows two examples of serum mass spectrum data, which are from an ovarian cancer and a control case respectively.

B. Preprocessing

The following conceptual model could consider for mass spectral data [7]. Suppose we observed N spectra, each taken on the same equally-spaced grid of length T of TOFs t_j, j=1, …, T. A model for $y_i(t_j)$, the observed spectral intensity for spectrum i at TOFs t_j, is

$$y_i(t_j) = B_i(t_j) + N_i S_i(t_j) + \varepsilon_{ij} \quad (1)$$

The true signal, $S_i(t)$, consist of a sum of possible peaks, each corresponding to a particular biological molecule. The normalization factor, N_i, is a constant multiplicative factor to adjust for spectrum-specific variability. The baseline function, B_i, represents a systematic artifact commonly seen in mass spectrometry data. The electrical noise, ε_{ij}, we assume that is zero-mean Gaussians with the variance a function of time.

TABLE I
DISTRIBUTION OF MASS SPECTRA DATA

Datasets	Num. of Cancer	Num. of Control	Num. of Benign
Ovarian 4-3-02	100	100	16
Ovarian 8-7-02	162	91	0

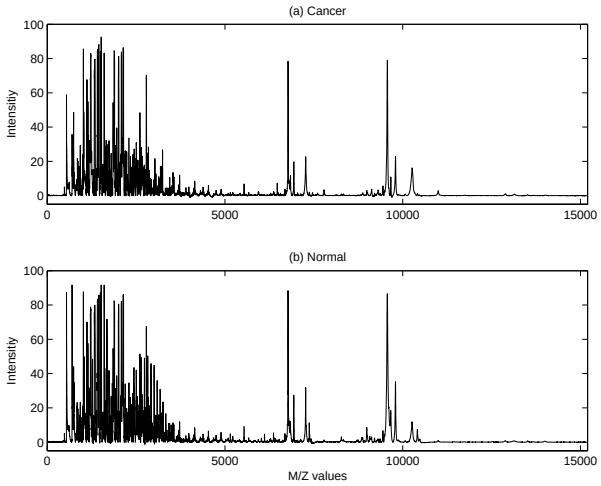

Fig. 1. SELDI-TOF MS sample plots. (a) an ovarian and (b) a control case.

According to above model, certain preprocessing steps must be performed before analyzing the spectra, including removal of baseline, noise elimination and normalization to calibrate the spectra from different samples. We performed baseline correction on all spectra by using a nonlinear filter known as the "top-hat" procedure [10]. The normalization is done via total ion current (TIC) method which is equivalent to the normalization with the L_1 norm of spectrum. Noise removal was done instead on wavelet coefficients. Fig. 2 shows a typical mass spectrum with the baseline and the same processed spectrum without the baseline.

III. Methods

A. Feature Extraction

As described in the previous sections, a mass spectrum is very spiky functional data that the key features are the peaks in the mass spectra. Hence, the analysis method must be considering both local and global behavior of the mass spectra and extracting the signal-pattern features. This fact encourages using a feature extraction method such as wavelets to account for this potentially useful information.

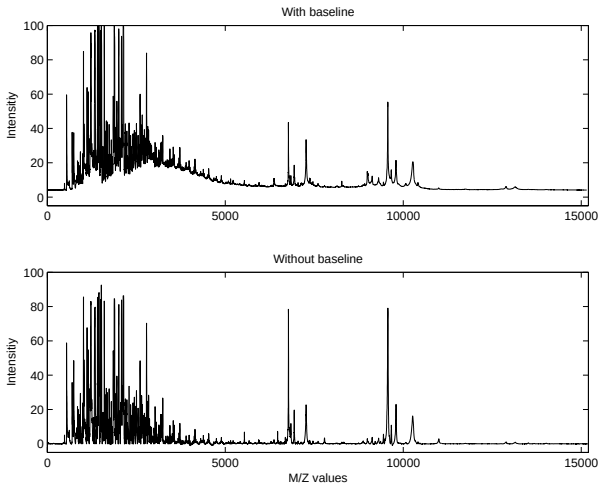

Fig. 2. A typical mass spectrum with and without the baseline.

Wavelets are families of orthonormal basis functions that can be used to parsimoniously represent other functions. For example, in $L^2(\Re)$, an orthogonal wavelet basis is obtained by dilating and translating a *mother wavelet* Ψ as $\Psi_{jk}(x) = 2^{j/2}\Psi(2^j x - k)$ with j, k being integers. A function f can then be represented by the wavelet series, as follows

$$f(x) = \langle f, \Phi \rangle \Phi(x) + \sum_{j=0}^{J} \sum_{k=0}^{2^j - 1} \langle f, \Psi_{j,k} \rangle \Psi_{j,k}(x) \qquad (2)$$

The set $\{\langle f, \Phi \rangle, \langle f, \Psi_{j,k} \rangle\}$ for $j = 0, \cdots, 2^J$, $k = 0, \cdots, 2^j - 1$ is the set of wavelet coefficients. As we ascend in the level of detail, increasing j, wavelet coefficients become smaller and smaller except from parts of the signal where spiky behavior is observed. By thresholding the wavelet coefficients we can reconstruct a de-noised version of the signal retaining the regions in which peaks are present [11].

B. Feature Selection

The feature selection methods are generally classified into two categories: filter and wrapper methods. In filter methods, the feature selector is independent of the specific learning algorithm used in classification and is used as a filter to discard irrelevant and/or redundant features. On the other hand, in wrapper methods, the feature selector behaves as a wrapper around the specific learning algorithm depending on which relevant features are determined [12].

In this study, we used a filter approach to feature selection in the mapped space of mass spectrum data. To do this, we implemented a statistical testing using a distance measure, called "Fisher's criteria," defined as follows

$$\mathrm{FC} = \frac{(\mu_1 - \mu_2)^2}{{\sigma_1}^2 + {\sigma_2}^2} \qquad (3)$$

Where μ_1 and μ_2 are the arithmetic means for the wavelet coefficients at each point of the cancer and non-cancer groups, respectively. σ_1 And σ_2 are the standard deviation of the corresponding coefficients at each point for both cancer and non-cancer groups.

IV. Results

We applied our approach for ovarian cancer detection using serum SELDI-TOF MS data, as described in Table I. As wavelet basis, we chose Daubechies, which has been reported previously to have a good performance on this field [13]. As a result, we considered an appropriate distance window to adequately differentiate various peaks corresponding to different molecules between selected points, which prevents the choice of points with correlated values [14][15].

We used three standard measures of the effectiveness of diagnosis technique: sensitivity, specificity, and accuracy. The *sensitivity* is the probability of the correct diagnosis for a patient with cancer, the *specificity* is the chances of the correct diagnosis for healthy person, and the *accuracy* is the probability of the correct diagnosis for the overall population of healthy and sick people. For evaluation of performance of our approach, we then applied linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and neural networks to the reconstructed mass spectra of ovarian cancer patients and healthy people.

We obtained wavelet transform of each mass spectrum in level, $J = 10$, with Daubechies four mother wavelet. Using Fisher's criteria, the twenty coefficients were selected from each approximation and details in the wavelet space. The other coefficients were essentially zero. Then, the processed mass spectrum of each sample was reconstructed from residual coefficients. Fig. 3 shows two examples of reconstructed mass spectrum, which are from a healthy and an ovarian cancer case respectively.

After reconstruction, the processed mass spectrometry raw data used to determine the differences between the ovarian cancer and healthy people cases. Because this is detection task, the processed serum samples in each dataset were divided into cancer and non-cancer groups in which the control and benign samples were grouped as a control set. Then, we split each dataset into training and testing using random permuting method. Half of samples in each group were used for the training and the remaining for the testing. Again, we applied the feature selection method to processed mass spectra for extraction of significant M/Z points. We implemented an experimental procedure that allowed us to control the number of M/Z points and minimal distance between selected points. We determined the number of points and minimal distance between points that lead to the highest accuracy in two datasets. The optimal number of selected points was determined to be ten and the minimum distance thirty-two.

Table II shows the results for correctly classified mass spectra from the testing set for each dataset described in Table I. For each classification, the accuracy, specificity, and sensitivity calculated and the results of detection on two datasets with three different classifiers are listed in Table III. The results show that all three classification

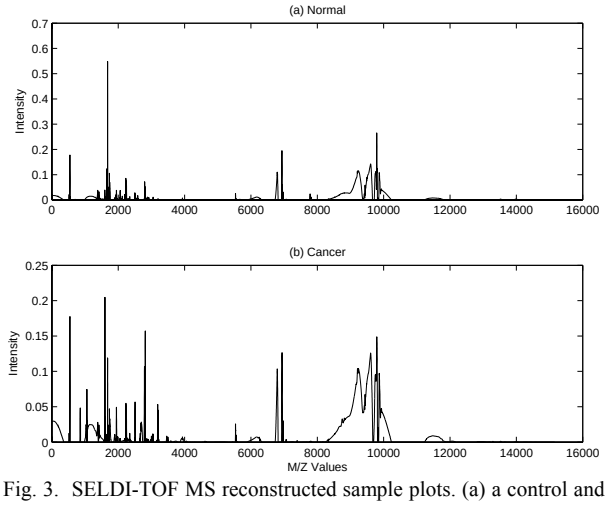

Fig. 3. SELDI-TOF MS reconstructed sample plots. (a) a control and (b) an ovarian case.

techniques reach the same accuracy, specificity, and sensitivity for each dataset with selected M/Z points. However, we have shown that the selected biomarkers could discriminate fairly the cancer case from non-cancer patients using proteomic patterns.

We identify a set of 10 M/Z ratios by our approach for the wavelet-based reconstructed mass spectra that the peptides are (432.3, 422.3, 718.9, 413.9, 407.5, 748.2, 236.2, 582.9, 244, and 958.6). Our results agree with the results obtained by other previously published works [13]. We have shown that the wavelet-based reconstructed mass spectrum is a viable approach which can be used to diagnosis of ovarian cancer from proteomic patterns.

V. Conclusions

We studied the problem of diagnosing ovarian cancer by analysis of mass spectrum data. We applied a wavelet-based preprocessing method to available datasets. We used a filter approach for selection of coefficients in the wavelet space, and reconstructed the processed mass spectra from the chose coefficients. We applied LDA, QDA, and neural networks (N.N.) to determine the effectiveness of these points in ovarian cancer detection. For available datasets, performance of our approach with selected points yielded an accuracy of 98.6%, specificity of 97.1%, and sensitivity of 100% for these classification techniques.

It must be pointed out that at the proteomic level; there may be two types of biomarkers that can be related to cancer. It could be that cancer results in the presence of specific proteins, which are not present in the non-cancer cases. The cancer can be diagnosed by detecting the physical presence of these specific proteins. Alternatively, the cancer may not lead to expression of a novel protein. Instead, it may change the complex proteomic pattern of the tumor-host microenvironment. In this case, the biomarker may be those normal host proteins that are aberrantly increased or decreased in abundance. This is an application where the feature extraction/selection techniques can be most helpful.

References

[1] E.F. Petricoin III, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *The Lancet*, Vol. 359, pp. 572-577, 2002.

[2] H. Kuruma, S. Egawa, M. Oh-Ishi, Y. Kodera, T. Maeda, "Proteome analysis of prostate cancer," *Prostate Cancer and Prostatic Diseases*, Vol. 8, pp. 14-21, 2005.

[3] T.P. Conrads, M. Zhou, E.F. Petricoin III, L. Liotta, T.D. Veenstra, "Cancer diagnosis using proteomics patterns," *Expert Rev. Mol. Diagn*., Vol. 3, No. 4, pp. 411-420, 2003.

[4] E.F. Petricoin III, D.K. Ornstein, C.P. Paweletz, A.M. Ardekani, P.S. Hackett, B.A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C.B. Simone, P.J. Levine, W.M. Linehan, M.R. Emmert-Buck, S.M. Steinberg, E.C. Kohn, L.A. Liotta, "Serum proteomic patterns for detection of prostate cancer," *Journal of National Cancer Institute*, Vol. 94, No. 20, pp. 1576-1578, October 2002.

[5] E.J. Finehout, K.H. Lee, "An introduction to mass spectrometry appli- cations in biological research," *Biochemistry and Molecular Biology Education*, Vol. 32, No. 2, pp. 93-100, 2004.

[6] A. Jemal, A. Thomas, T. Murray, M. Thun, "Cancer statistics," *CA Cancer J. Clin*., Vol. 52, pp. 23-47, 2002.

[7] J.S. Morris, K.R. Coombes, J. Koomen, K.A. Baggerly, R. Kobayashi, "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum," *Bioinformatics*, Vol. 21, No. 9, pp. 1764-1775, 2005.

[8] S. Mallat, "A wavelet tour of signal processing," *Academic press*, 1998.

[9] K.A. Baggerly, J.S. Morris, K.R. Coombes, "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics*, Vol. 20, No. 5, pp. 777-785, 2004.

[10] A.G. Hanbury, J. Serra, "Morphological operators on the unit circle," *IEEE Trans. Image Processing*, Vol. 10, No. 12, pp. 1842-1850, 2001.

[11] D.L. Donoho, "De-noising by soft thresholding," *IEEE Trans. Information Theory* , Vol. 41, pp. 613-627, 1995.

[12] G. Frosini, B. Lazzerini, F. Marcelloni, "A modified fuzzy C-means algorithm for feature selection," *IEEE*, 2000.

[13] M. Vannucci, N. Sha, P.J. Brown, "NIR and mass spectra classification: Bayesian methods for wavelet-based feature selection," *Chemometrics and Intelligent Laboratory Systems*, Vol. 77, pp. 139-148, 2005.

[14] H. Tong, Y. Mukomel, E. Fink, "Diagnosis of ovarian cancer based on mass spectra of blood samples," *IEEE*, 2004.

[15] L. Li, H. Tang, Z. Wu, J. Gong, M. Gruidl, J. Zou, M. Tockman, R.A. Clark, "Data mining techniques for cancer detection using serum proteomic profiling," *Artificial Intelligence in Medicine*, Vol. 32, pp. 71-83, March 2004.

TABLE II
CLASSIFICATION OF SERUM SAMPLES

Datasets		Cancer	Control
Ovarian 4-3-02	Cancer	50/50	2/58
	Control	0/50	56/58
Ovarian 8-7-02	Cancer	81/81	1/46
	Control	0/81	45/46

TABLE III
PERFORMANCE OF CLASSIFICATION OF TEST SAMPLES

Datasets		LDA	QDA	N.N.
Ovarian 4-3-02	Accuracy	98.1%	98.1%	98.1%
	Sensitivity	100%	100%	100%
	Specificity	96.5%	96.5%	96.5%
Ovarian 8-7-02	Accuracy	99.2%	99.2%	99.2%
	Sensitivity	100%	100%	100%
	Specificity	97.8%	97.8%	97.8%